基于 GNN 的文本分类算法研究

高　贵,赵　阳,于舒娟,姚成杰,黄丽亚
(南京邮电大学 电子与光学工程学院,江苏 南京 210046)

摘　要:图神经网络(Graph Neural Networks,GNN)因其结构的新颖性在文本分类任务中广受关注。针对 GNN 在训练数据集较少时容易出现过拟合、特征信息不足等问题,提出了 Att-DASA-ReGNN(Regional Embedding GNN based on Data Augmentation and Self-Attention with the Attention Mechanisms)模型。该模型在数据特征提取阶段引入了简单数据增强方法(Easy Data Augmentation,EDA)和 Self-Attention 技术改善了过拟合问题;原模型词嵌入方式对维度很高且稀疏的高阶邻域信息的捕捉能力不足,该模型中通过增加区域词嵌入技术,加强了词级之间的关系,使得模型更容易捕捉高阶邻域信息,从而减轻数据稀疏带来的影响。为了进一步提升模型的文本分类准确率,该模型的图词特征交互阶段通过引入 Soft-Attention 技术改进了注意力权重提取方式。最后,在多种数据集上的实验证明,该模型的分类准确率较之前模型均有不同程度的提升。

关键词:图神经网络;文本分类;数据增强;词嵌入;注意力机制
中图分类号:TP391.1　　**文献标识码**:A　　**文章编号**:1673-629X(2023)05-0138-07
doi:10.3969/j.issn.1673-629X.2023.05.021

Research on Text Classification Algorithm Based on GNN

GAO Gui,ZHAO Yang,YU Shu-juan,YAO Cheng-jie,HUANG Li-ya
(School of Electronic and Optical Engineering,Nanjing University of Posts and Telecommunications,
Nanjing 210046,China)

Abstract:Graph Neural Networks (GNN) has gained wide attention in text classification tasks because of its novelty of structure. Aiming at the problems that GNN is prone to overfitting and insufficient feature information when the training dataset is small,we propose the Att-DASA-ReGNN model. The model applied EDA and Self-Attention techniques to improve the overfitting in the data feature extraction stage. Aiming at the problem that the word embedding method of the original model is insufficient to capture high-order neighborhood information with high dimensions and sparseness,the regional word embedding technology is introduced in the model. The technique further strengthens the relationship between word levels,making it easier for models to capture high-order neighborhood information,thereby mitigating the impact of data sparsity. In order to further improve the text classification accuracy of the model,the interactive stage of the graph-word features of the model improves the attention weight extraction method by introducing Soft-Attention technology. Finally,simulation experiments on various datasets show that the classification accuracy of the model is improved to varying degrees compared with the previous model.

Key words:graph neural network;text classification;data enhancement;word embedding;attention mechanism

0　引　言

在信息数字化的 21 世纪,自然语言处理(Natural Language Processing,NLP)在人工智能研究中的地位越来越重要。作为 NLP 领域的重要分支之一,文本分类技术常被用于处理复杂多样的文本信息,其主要工作是根据特征对文本进行分类,并为其分配不同的标签。基于文本分类技术,用户可以通过搜索关键词或查找相应标签,快速准确地找到所需信息。

在基于深度学习的文本分类技术中,图神经网络(Graph Neural Networks,GNN)对非欧几里德数据的独特建模方式吸引了学者们的广泛研究。2004 年,Mihalcea 等人[1]首次将图模型应用于文本分类任务,TextRank 通过图论将自然语言中的文本重新定义表示,包括单词、短语、完整句子等。Defferrard 等人[2]提

收稿日期:2022-07-08　　**修回日期**:2022-11-10
基金项目:国家自然科学基金(61871234)
作者简介:高　贵(1998-),男,硕士,通讯作者,研究方向为智能化信号处理及大数据分析;于舒娟(1967-),女,硕士,副教授,研究方向为智能化信号处理及大数据分析。

出了基于图的卷积神经网络模型 Graph-CNN,首次将文本转化为一组词的图的集合,利用图卷积对每个子图进行卷积运算。Yao 等人[3]构建了一个简单高效的图卷积网络 TextGCN,模型基于词的共现性和词与词之间的相互关系,将待分类的整个文本数据集构建成文本图。该方法考虑了节点的高阶邻域信息,有效提高了文本分类性能。

考虑到要进一步增强依赖关系的捕捉能力和提高计算效率,Vaswani 等人[4]提出了自注意力机制,并在 NLP 领域中广泛应用。Veličković等人[5]提出了图注意网络,成功地将带有掩码的自注意力机制与 GNN 结合。Zhang 等人[6]提出了一种基于自注意力机制的 GNN 模型,该网络适用于具有同构或异构的超图,有助于图形表示学习,可以表示不同应用中复杂的高阶相互作用。

然而,虽然这些图神经网络和注意力机制方法在文本分类任务上取得了成功,但是在图神经网络的数据增强以及关键信息的权重计算方面还应用较少。在研究中发现,由于 GNN 的特殊图文转换特征,其在训练数据集较少时容易出现过拟合的问题。此外,由于传统 GNN 模型的词嵌入方式对高阶领域信息捕捉能力不足,当训练数据稀疏时会对模型性能带来负面影响。

基于以上 GNN 模型在文本分类任务中所遇到的问题,该文提出了 Att-DASA-ReGNN 模型。该模型主要有如下三点创新改进:

(1)针对模型训练中容易出现的过拟合问题,在模型的数据特征提取阶段应用了 EDA 技术和 Self-Attention 技术。该技术在扩充数据集的同时加强了单词级别的相互联系,改善了过拟合问题。

(2)针对原模型词嵌入方式对维度很高且稀疏情况下的高阶邻域信息捕捉能力不足的问题,在模型中引入了区域词嵌入技术。该技术进一步加强了词级之间的关系,使得模型更容易捕捉高阶邻域信息,从而减轻数据稀疏带来的影响。

(3)为了进一步提升模型的文本分类准确率,在模型的图词特征交互阶段改进了注意力权重提取方式。通过引入三种不同的注意力机制验证模型性能的提升效果,最终确定为 Soft-Attention 作为该阶段的注意力权重提取方式。

1　相关工作

前馈神经网络是最早用于文本分类的深度学习模型。它们使用词嵌入模型来学习文档中文本表示,将文本中的词向量相加的和或平均值作为输出将其送入前馈神经网络中[7]。2015 年,Zhou 等人结合 CNN 和 RNN 两者的优势,提出了一个 C-LSTM 模型[8]。该模型首先利用 CNN 提取高层次的特征,然后将特征送入 LSTM 以获得句子表示。2017 年,王俊丽等人提出了一个 ResLCNN 模型[9]。该模型不仅将 LSTM 与 CNN 结合起来提取更复杂的抽象特征,而且还使用残差来缓解 LSTM 梯度消失的问题。2018 年,谭咏梅等人利用卷积神经网络结合双向 LSTM 从文本中提取特征[10]。该模型将得到的特征输入全连接层,然后利用语义规则进一步处理分类结果,最终提高了中文文本的分类性能。

图神经网络以其在分类精度的优越性,被广大研究学者应用于文本分类领域。Bruna 等人将欧几里德空间卷积转移到图网络中,并为谱域和空间域提出了相应的图卷积方法[11]。Henaff 等人将图卷积应用于神经网络,对有和无输入标签的大型数据集都进行分类[12]。Defferrard 等人[2]在图谱域定义并应用卷积,解决了 Bruna 等人的计算高复杂性和滤波器的非局部问题。Li 等人提出了一种能够处理任何图结构的图网络,以解决以前的图卷积神经网络面临的固定滤波器和图结构的问题[13]。Huang 等人重新改进了图神经网络的结构,将单个文本视作图,用词共现方法构建词之间的关系,最后用图卷积神经网络提取特征,在提升模型性能的同时还减少了不必要的内存消耗[14]。Zhang 等人使用门控图神经网络提出了一种基于 GNN 的归纳式文本分类方法,同时提出了不同的构建文本图的方法[15]。该方法通过训练样本获得词之间的相互关系,该模型对于有较多新词的文本分类数据集效果更好。

数据采样部分如图 1 所示,简单数据增强技术 EDA 是 Wei 等人提出来的一种数据增强方法,包含了四种类型,分别是:

(1)同义词替换:在一个句子中随机抽取其中的词,用这些词的近似词进行同义替换,形成新的句子。

(2)随机插入:在一个句子中随机选择一个词,之后用该词的同义词随机插入该句子中的任意位置。

(3)随机交换:将一个句子中任意选定的两个单词进行互换位置。

(4)随机删除:将一个句子中的任意单词以概率 p 进行概率性随机删除。

使用 EDA 对文本进行数据增强后,可以得到数倍于原数据的有效数据。

接着,Att-DASA-ReGNN 模型在 EDA 数据增强后引入了自注意力机制 Self-attention。这样做的目的是将增强后的数据集通过两个神经网络层和一个归一化层组成的模块,让提取到的特征拥有更多的细节。自注意力的计算公式可以表示为:

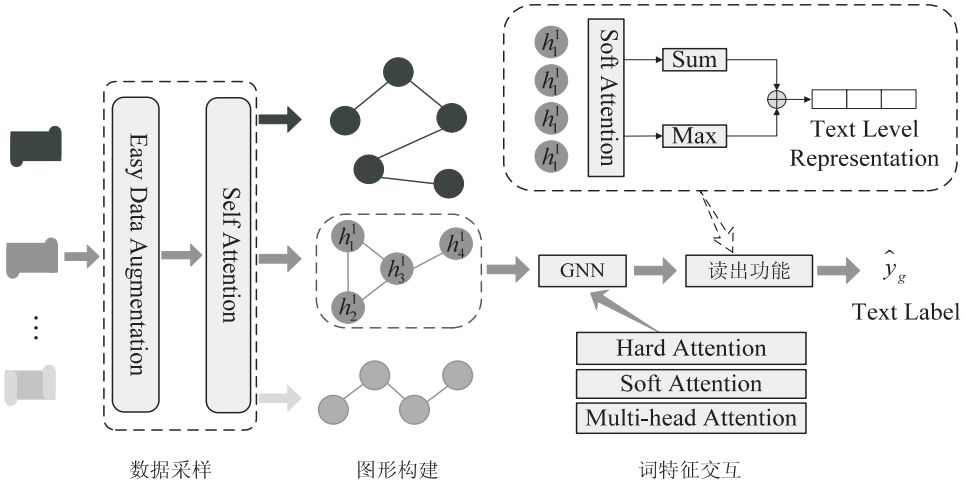


图 1 Att-DASA-ReGNN 模型

$$\mathrm{Att}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \omega(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}})\boldsymbol{V} \tag{1}$$

其中，$\boldsymbol{Q}$ 是查询向量矩阵，$\boldsymbol{K}$ 是键向量矩阵，$\boldsymbol{V}$ 是值向量矩阵。

如图 2 所示，$\boldsymbol{X}_i$ 为词嵌入产生的词向量。接着词向量 $\boldsymbol{X}_i$ 分别与三个矩阵 $\boldsymbol{W}^{(q)}$、$\boldsymbol{W}^{(k)}$、$\boldsymbol{W}^{(v)}$ 相乘得到三个矩阵向量 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$。每一个 $\boldsymbol{Q}_i$ 与所有的 $\boldsymbol{K}_i$ 进行矩阵乘法得到 α_{ij}，其中 $\boldsymbol{Q}_i$ 与 $\boldsymbol{K}_i$ 进行相乘之后需要除一个 d，d 是 $\boldsymbol{Q}_i$ 与 $\boldsymbol{K}_i$ 的维度。最后，每一个 α_{ij} 经过 SoftMax 层之后得到了 β_{ij}，之后将所有的 β_{ij} 相加即可得到输出 b_1，即词 $\boldsymbol{X}_i$ 的自注意力机制得分。

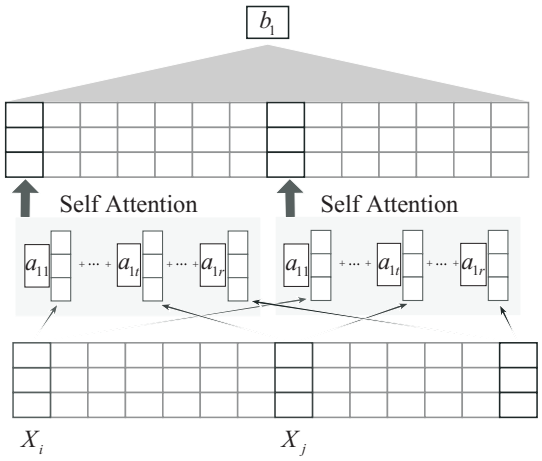


图 2 自注意力机制流程示意图

区域词嵌入是专注于学习文本区域特征的词嵌入方法。该方法在进行区域特征表示的同时保留了原本数据集的内部结构信息。其中，区域可以理解为文本中固定长度的连续子序列，用 w_i 表示句子中的第 i 个词，用 region(i,c) 表示当前第 i 个词与该词前后一共 $2c+1$ 个词组成的短语。Att-DASA-ReGNN 模型中，区域词嵌入方式用 e_w 表示第 w 个词的嵌入，该嵌入可以用矩阵 $\boldsymbol{E} \in R^{h\times v}$ 表示，其中 v 表示词汇的大小，h 表示嵌入的大小。区域词嵌入的具体流程如图 3 所示。

为了利用单词的相对位置和本地语境的信息，除了学习单词嵌入外，还为每个单词学习了一个局部的语境单元，表示为矩阵 $\boldsymbol{K}_{w_i} \in R^{h\times(2c+1)}$，$\boldsymbol{K}_{w_i}$ 中的每一列都可以用来与相应的 w_i 进行相对位置上的上下文词的交互。

事实上，单位矩阵的列可以被视为独特的线性映射函数，这些映射函数可以进行学习来捕捉语义进行词的嵌入，用 $p^i_{w_{i+t}}$ 表示 w_{i+t} 在第 i 个词的映射词嵌入，表示为式(2)。

$$p^i_{w_{i+t}} = K_{w_i,t} \odot e_{w_{i+t}} \tag{2}$$

2 Att-DASA-ReGNN 模型

Att-DASA-ReGNN 主要由四个部分组成：第一部分是自注意力机制 Self attention 和 EDA 结合生成的数据增强数据采样部分；第二部分是利用滑动窗口进行图形构建；第三部分为基于门控图神经网络(Gated Graph Neural Network，GGNN)的词特征交互；最后将提取到的特征送入两个多层感知机(Multi-Layer Perceptron，MLP)完成文本的预测分类。图 1 为 DASA-GNN 模型的结构框图。

2.1 图形构建和词特征交互

如图 1 所示，首先将句子中选中的单词表示为节点，接着用单词之间的共现形式表示为边来进行图形构建，图可以用 $G=(V,E)$ 表示，其 V 表示图形的节点，E 表示图形的边。共现指的是在滑动窗口中单词的相关性，其中滑动窗口大小一般默认设定为 3，其中的边都为无向边。Nikolentzos 等人[16]将滑动窗口的大小定义为 2。他们将图视为密集连接的图，其模型中图消息的传递机制主要是用一个特定的基本节点与其他每一个节点相连，因此在该图中只能得到模糊的结构消息。而门控图神经网络 GGNN 中为避免图的密集连接导致的单词特征信息模糊，会首先初始化文本数据的词特征来进行节点的嵌入表示，接着将任意一个文档都进行了单独子图表示，因此在模型中词交互

阶段部分,单词特征信息能够清晰地传播到上下文中[17]。

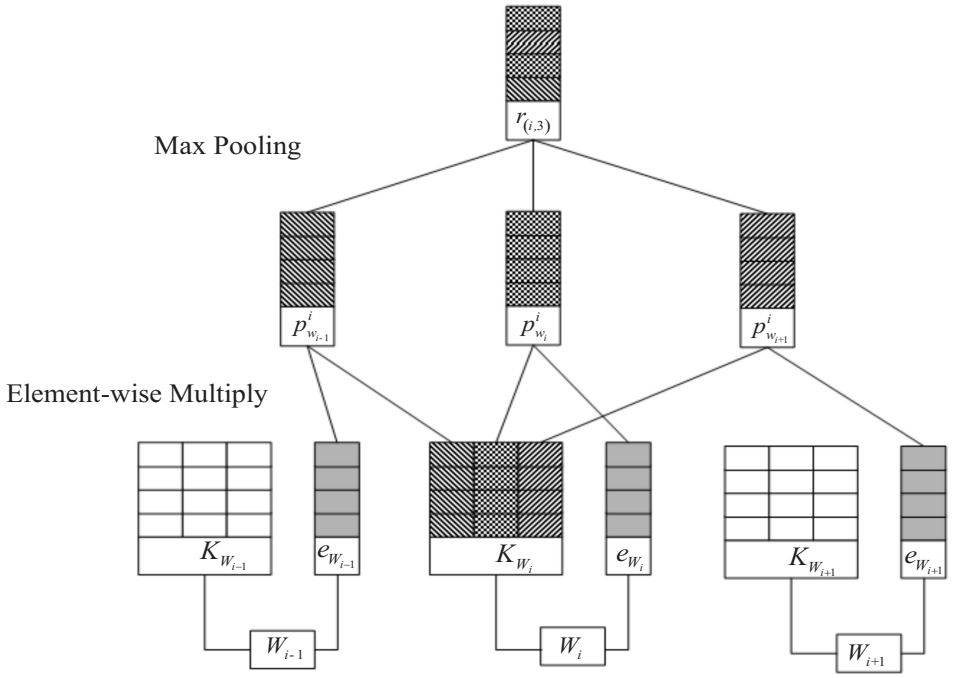


图 3　区域词嵌入流程

2.2　特征交互的注意力权重提取方式

为了进一步提升 Att-DASA-ReGNN 模型的准确率,从模型的图词特征交互的角度出发,在特征交互的注意力权重提取中分别引入了硬注意力、软注意力和多头注意力机制。

硬注意力机制特点是在模型中引入了位置变量 s_t ,它表示在模型生成第 t 个词语时注意到的图片位置。其中,s_t 中第 i 个分量记为 $s_{t,i}$,其取值只能为 0 或者 1。当图像中第 i 个子区域被模型用于特征抽取的时候,则 $s_{t,i}=1$,反之,则 $s_{t,i}=0$。硬注意力机制规定在 l 个分量中只能有一个权值为 1,其余值必须全为零。因此可以理解为该向量 s_t 是 One-hot 形式的向量。这一规定意味着硬注意力机制每次只在图像中的 l 个区域中随机选取一个区域并加以特别关注,而这也是其名称的由来。此外,硬注意力机制将 $s_{t,i}$ 视为一个中间的隐变量,与此同时令 $s_{t,i}$ 服从参数量为 $\{\alpha_{t,i}\}$ 多元伯努利分布。最后,硬注意力机制还将 $\hat{z}_t$ 视为一个随机变量,即有式(3)和式(4)。

$$p(s_{t,i}=1 \mid s_{j<t},a_i)=\alpha_{t,i} \tag{3}$$

$$\hat{z}_t=\sum_i s_{t,i}a_i \tag{4}$$

在分类问题中,经常被提到的就是软注意力机制。其主要思想是,首先将 Source 中的构成元素想象成是由一系列的<Key,Value>数据对构成,此时给定 Target 中的某个元素 Query,通过计算 Query 和各个 Key 的相似性或者相关性,得到每个 Key 对应 Value 的权重系数,然后对 Value 进行加权求和,得到了最终需要的注意力数值。所以,本质上软注意力机制是对 Source 中元素的 ValueValue 值进行加权求和,而 Query 和 KeyKey 用来计算对应 Value 的权重系数。即可以将其本质思想表示为式(5)。

$$\text{Attention}(\text{Query},\text{Source})=\sum_{i=1}^{L_x\sum(\text{Query},\ \text{Key}_i)\text{Value}_i}\text{Similarity} \tag{5}$$

多头注意力机制是对注意力机制的每个头进行运算,是对于输入 Query、Key、Value 进行的运算,然后把每个头的输出拼起来乘以一个矩阵进行线性变换,得到最终的输出,其表达式为式(6)。

$$h_i^{(l+1)}=\sigma\left(\frac{1}{k}\sum_{k=1}^{K}\sum_{j\in N(i)}\alpha_{ij}^k\boldsymbol{W}^k h_j^{(l)}\right) \tag{6}$$

其中,K 表示多头注意力机制的个数,α_{ij}^k 是由第 K 个头注意力机制的归一化参数,$\boldsymbol{W}^k$ 是线性变换权重矩阵。

3　实　验

本节将在多个不同的数据集上进行一系列实验,验证所提模型的文本分类性能,而且为更精确地展现 Att-DASA-ReGNN 模型的分类性能,选取了文本分类方面的几种经典算法模型以及最新的研究成果模型作为实验的对照组。

3.1　实验数据集

为了验证模型的性能及其稳定性,挑选 5 种不同的英文数据集来比较模型的分类效果。这些数据集是:

(1)MR:MR 数据集属于电影评论领域。它是一个二分类数据集,其中每个评论仅包含一句话,分为正面评论和负面评论。

(2)R8:R8 数据集属于新闻领域。它是从路透社的新闻专线中收集分类得到的,总共分为 8 类。

(3)SST1:SST1 数据集属于社会领域。它来自于斯坦福情感树库,包括非常消极、消极、中性、积极、非常积极五种类型的数据。

(4)SST2 数据集与 SST1 数据集相同,但去掉了中性评论和二进制标签,只保留了两类标签。

(5)SUBJ 数据集是主观性数据集,该数据集用主客观的指标将句子进行二分类。

(6)TREC 数据集为问句类型的数据集。

这些实验数据集的详细信息如表 1 所示。

表 1 实验数据集详细信息

数据集	类别数	平均句子长度	最大句子长度	训练集	测试集	文档大小
MR	2	20	56	9 595	1 050	10 662
R8	8	41	291	5 485	2 189	7 674
SST1	5	18	53	9 645	2 210	11 855
SST2	2	19	53	7 792	1 821	9 613
SUBJ	2	23	120	9 000	1 000	10 000
TREC	6	10	37	5 452	500	9 592

3.2 对比模型

为更精确地展现 DASA-GNN 模型的分类性能,本章选取了文本分类方面的几种经典算法模型以及最新的研究成果模型作为实验的对照组。所选模型按照原理大致可以分为深层神经网络模型和基于图的网络模型,具体介绍如下:

(1)CNN(non-static):该模型将卷积神经网络应用于文本分类,并使用了随机初始化单词嵌入来提取句子的关键信息。

(2)CNN(rand):该模型同样基于卷积神经网络,与 CNN(non-static)不同的是,它使用了预训练单词嵌入来提取句子关键信息。

(3)BiLSTM(RNN):该模型使用双向 LSTM 结构进行文本分类,并使用了预训练单词嵌入提取信息。

(4)Texting(GNN):Texting 为每个文档构成单独的图,并利用 GGNN 进行文本分类。

(5)TextGCN(GCN):TextGCN 将整个语料库构成一个图,并应用 GCN 进行文本分类。

3.3 评价指标

为评价文中改进模型对文本分类的有效性,采用准确率作为评价指标。其公式可以表示为:

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{7}$$

其中,TP 表示真正例,TN 表示真反例,FP 表示假正例,FN 表示假反例。

3.4 实验结果

为了对 DASA-GNN 模型进行全面分析,本节设计了多组实验。实验 1:基于自注意力机制的 EDA 的数据增强效果验证实验;实验 2:基于 Dropedge 技术的区域词嵌入对模型性能提升效果验证实验;实验 3:三种不同的注意力权重提取方式对模型性能提升效果验证实验;实验 4:模型超参数设置对模型分类性能的影响实验。具体分析如下:

(1)为了更好地了解 EDA 数据增强的作用及其对 DASA-GNN 模型的性能影响,实验 1 选择了 Texting 作为对照模型,并选择了不同百分比训练数据下的准确率作为对比结果考量。实验结果如表 2 所示,对于 MR、SST2 和 SUBJ 数据集上的实验结果而言,DASA-GNN 模型的最佳性能对比 Texting 模型有 2.93 百分点、1.52 百分点和 0.15 百分点的提升,且最佳性能都在 30% 数据量时出现。对于 R8 和 SST1 数据集而言,服务器上得出的最佳结果尽管略微不如 Texting 模型,但是符合实验预期的结果。由此可见,在模型的数据增强部分加入自注意力层后可以进一步改善数据质量,提升模型性能。

(2)将引入区域词嵌入的 DASA-ReGNN 模型与七个深度学习领域的算法进行对比,最终对比结果见表 3。表中模型在不同数据集上的最佳准确率用加粗字体表示,次优准确率用下划线表示。从表 3 中可以看出,DASA-ReGNN 模型在多个数据集上都表现出了优异的性能。其中,在 R8、SUBJ 数据集上显示图神经网络具有良好的分类性能,而 DASA-ReGNN 通过引入区域词嵌入表现得更为优异;与其余六个深度学习领域的经典以及最新的算法的最佳性能相比,DASA-ReGNN 还提升了 0.36 百分点和 0.24 百分点的分类精度。在除 SST1 以外的其余五个数据集上 DASA-ReGNN 都提升了一定的分类精度,表现出了良好的模型性能。

表 2　引入自注意力机制的模型性能比较

数据量百分比/%	MR	R8	SST1	SST2	SUBJ
10	0.682 6	0.825 5	0.404 1	0.819 2	0.901 1
20	0.740 4	0.887 3	0.447 3	0.825 1	0.902 8
30	**0.827 5**	0.967 4	0.451 6	**0.882 6**	**0.941 2**
40	0.813 2	0.968 5	0.447 9	0.860 4	0.935 5
Texting	0.798 2	**0.980 4**	**0.461 2**	0.867 4	0.939 7

表 3　引入区域词嵌入的 GNN 与其他网络模型的准确率比较

Model	MR	R8	SST1	SST2	SUBJ	TREC
CNN(rand)	0.761 0	0.940 2	0.450 0	0.827 0	0.896 0	0.912 0
CNN(non-static)	0.815 0	0.957 1	**0.480 0**	0.872 0	0.934 0	0.936 0
TextCNN	0.773 3	-	0.457 9	0.868 1	0.905 0	0.922 0
RNN(BiLSTM)	0.776 8	0.963 1	0.459 0	0.858 0	0.916 0	0.910 0
TextGCN	0.767 4	0.970 7	-	-	-	-
GNN(Texting)	0.798 2	0.980 4	0.461 2	0.867 4	0.939 7	0.933 4
DASA-ReGNN	**0.829 4**	**0.984 0**	0.453 5	**0.887 6**	**0.942 1**	**0.952 0**

(3)实验中采用了三种不同的注意力权重提取方式对模型性能提升效果进行验证,其结果如表 4 所示。从四个模型在不同数据集上的分类结果对比可以得出,在 Att-DASA-GNN 模型中引入不同的注意力机制可以有效提升文本分类的性能。例如在 MR 数据集上,DASA-ReGNN 的分类准确率为 0.829 4,而引入硬注意力机制的 Att-DASA-ReGNN 模型的准确率可达 0.830 0,引入软注意力机制和多头注意力机制的模型准确率提升效果更好,分别为 0.841 0 和 0.832 4。在其他数据集上的提升效果也较明显。

表 4　不同注意力权重提取方式效果对比

Model	MR	SST2	SUBJ	TREC
DASA-ReGNN	0.829 4	0.887 6	0.942 1	0.952 0
DASA-ReGNN+Hard	0.830 0	0.887 9	0.942 7	0.954 2
DASA-ReGNN+Soft	**0.841 0**	**0.891 8**	**0.944 1**	**0.961 4**
DASA-ReGNN+Multihead	0.832 4	0.889 0	0.943 4	0.958 7

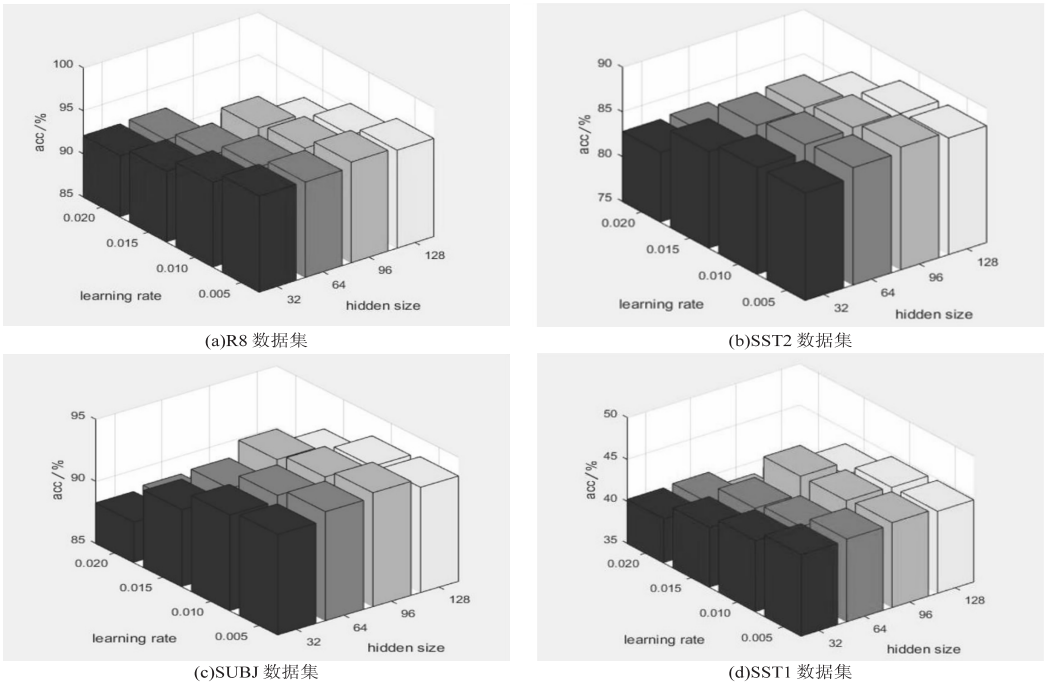
表 5 为三种不同注意力机制的 Att-DASA-ReGNN 模型与传统文本分类模型的性能比较实验结果。由表 5 可以得出,相对于一些传统模型,三种 Att-DASA-ReGNN 模型的文本分类准确率均有不同程度的提高,其中以软注意力机制模型 Att-DASA-ReGNN+Soft 的分类准确率最佳。例如,在 MR 数据中,性能表现最好的传统模型 CNN(non-static)的分类准确率为0.815 0,而三种 Att-DASA-ReGNN 模型的分类准确率分别为 0.830 0、0.841 0 和 0.832 4,均超过其他对照模型。此外,软注意力机制模型 Att-DASA-ReGNN+Soft 的分类性能在 4 个数据集上的分类准确率最高。由此可见,在模型图词特征交互中加入注意力机制的方法可以有效提升文本分类准确率,并且软注意力机制的提升效果最为有效。

表 5　Att-DASA-ReGNN 模型与其他的模型的分类准确率对比

Model	MR	SST2	SUBJ	TREC
CNN(rand)	0.761 0	0.827 0	0.896 0	0.912 0
CNN(non-static)	0.815 0	0.872 0	0.934 0	0.936 0
RNN(BiLSTM)	0.776 8	0.858 0	0.916 0	0.910 0
TextGCN	0.767 4	-	-	-
GNN(Texting)	0.798 2	0.867 4	0.939 7	0.933 4
Att-DASA-ReGNN+Hard	0.830 0	0.887 9	0.942 7	0.954 2
Att-DASA-ReGNN+Soft	**0.841 0**	**0.891 8**	**0.944 1**	**0.961 4**
Att-DASA-ReGNN+Mutihead	0.832 4	0.889 0	0.943 4	0.958 7

(4)最后,为探究两个重要超参数 learning rate 和 hidden size 对模型性能的影响,选择了在 R8、SST2、SUBJ 和 SST1 数据集上进行模型训练做进一步测试。实验结果如图 4 所示。

当 hidden size 参数不变时,DASA-GNN 模型的准确率在 learning rate 参数数值为 0.005 时达到最大;当 learning rate 参数不变时,hidden size 参数为 96 模型准确率达到最高。例如在 SUBJ 数据集上,从图 4(c)中可以看到,hidden size 参数为 96 的柱形图为模型准确率的最高值,并且其随着 learning rate 参数的提升而不断升高。由此可见,DASA-ReGNN 模型训练中的超参数 learning rate 和 hidden size 最优值依旧为 0.005 和 96,模型性能稳定可靠。



(a)R8 数据集 (b)SST2 数据集

(c)SUBJ 数据集 (d)SST1 数据集

图 4 Att-DASA-ReGNN 模型在不同 learning rate 和 hidden size 下的准确性比较

4 结束语

针对现有的基于图神经网络的文本分类方法存在的过拟合、特征稀疏和特征多样性不足等问题,提出了 Att-DASA-ReGNN。Att-DASA-ReGNN 模型在保留图神经网络中图形编码特性的同时,使用了基于自注意力机制的 EDA 数据增强技术,同时在图词特征交互阶段引入了区域词嵌入技术改善了高阶领域信息的捕捉问题,最后在图词特征交互阶段改进了注意力权重提取方式。实验表明,相较于其他现有模型,Att-DASA-ReGNN 模型在多个不同种类数据集上的分类准确率均有不同程度的提升,性能优越性显著。

参考文献:

[1] MIHALCEA R, TARAU P. Textrank: bringing order into text [C]//Conference on empirical methods in natural language processing. Barcelona: ACL, 2004: 404-411.

[2] DEFFERRARD M, BRESSON X, VANDERGHEYNST P. Convolutional neural networks on graphs with fast localized spectral filtering[J]. Advances in Neural Information Processing Systems, 2016, 29: 3837-3845.

[3] YAO L, MAO C, LUO Y. Graph convolutional networks for text classification [C]//AAAI conference on artificial intelligence. Honolulu: AAAI, 2019: 7370-7377.

[4] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.

[5] Veličković P, CUCURULL G, CASANOVA A, et al. Graph attention networks[J]. arXiv: 1710.10903, 2017.

[6] ZHANG R, ZOU Y, MA J. Hyper-SAGNN: a self-attention based graph neural network for hypergraphs[J]. arXiv: 1911.02613, 2019.

[7] 张 钺. 基于情感词向量优化的微博评论情感分析[D]. 长沙: 湖南大学, 2019.

[8] ZHOU C, SUN C, LIU Z, et al. A C-LSTM neural network for text classification[J]. Computer Science, 2015, 1(4): 39-44.

[9] 王俊丽, 杨亚星, 王小敏. 短文本分类的 ResLCNN 模型[J]. 软件学报, 2017, 28(2): 61-69.

[10] 谭咏梅, 刘姝雯, 吕学强. 基于 CNN 与双向 LSTM 的中文文本蕴含识别方法[J]. 中文信息学报, 2018, 32(7): 11-19.

[11] ESTRACH J B, ZAREMBA W, SZLAM A, et al. Spectral networks and deep locally connected networks on graphs[C]// International conference on learning representations. Canada: ICLR, 2014: 1-14.

[12] HENAFF M, BRUNA J, LECUN Y. Deep convolutional networks on graph-structured data[J]. arXiv: 1506.05163, 2015.

[13] LI R, WANG S, ZHU F, et al. Adaptive graph convolutional neural networks[C]//AAAI conference on artificial intelligence. New Orleans: AAAI, 2018: 3546-3553.

[14] HUANG L, MA D, LI S, et al. Text level graph neural network for text classification[J]. arXiv: 1910.02356, 2019.

[15] ZHANG Y, YU X, CUI Z, et al. Every document owns its structure: Inductive text classification via graph neural networks[J]. arXiv: 2004.13826, 2020.

[16] NIKOLENTZOS G, TIXIER A, VAZIRGIANNIS M. Message passing attention networks for document understanding[C]// Proceedings of the AAAI conference on artificial intelligence. New York: AAAI, 2020: 8544-8551.

[17] LI Y, TARLOW D, BROCKSCHMIDT M, et al. Gated graph sequence neural networks[J]. arXiv: 1511.05493, 2015.